Indirect and Suboptimal Control of Gene Expression is Widespread in Bacteria

Morgan N. Price[1,*], Adam M. Deutschbauer[1], Jeffrey M. Skerker[2,3], Kelly M. Wetmore[1,3], Troy Ruths[1], Jordan S. Mar[2,3], Jennifer V. Kuehl[1], Wenjun Shao[4], Adam P. Arkin[1,2,3,*]

1 Physical Biosciences Division, Lawrence Berkeley National Lab, 1 Cyclotron Road Mailstop 977-152, Berkeley California 94720, USA

2 Department of Bioengineering, University of California, Berkeley California 94720, USA

3 Energy Biosciences Institute, University of California, Berkeley California 94720, USA

4 Department of Molecular and Cell Biology, University of California, Berkeley California 94720, USA

∗ **E-mail:** morgannprice@yahoo.com, aparkin@lbl.gov

DRAFT July 16, 2012

Abstract

Gene regulation in bacteria is usually described as an adaptive response to an environmental change so that genes are expressed when they are required. However, the extent to which bacterial gene regulation is adaptive has not been tested systematically. By examining genome-wide data on mutant fitness and gene expression, we show that for diverse bacteria, there is little correlation between when genes are important for optimal growth or fitness and when they are more highly expressed. Genes that have a major impact on growth in a subset of conditions, such as biosynthetic or catabolic genes, show more of a correlation than other genes do, but even these genes are often not up-regulated when needed. Indeed, genes that are important for growth in the same conditions and that have closely-related functions often have little similarity in their expression patterns. Conversely, genes are often highly-expressed and detrimental to fitness instead of being repressed when not needed. Many genes are regulated by growth rate rather than by a functionally-relevant cue, while other functionally-diverse genes are expressed constitutively, regardless of need. To explain our data, we propose that indirect control, wherein the expression of a gene responds to signal(s) that are not related to the function of the gene, evolves more readily than direct control by a functionally-relevant signal. Indirect control is often maladaptive in artificial conditions, and we propose that it is adaptive but suboptimal in the wild.

Author Summary

Microorganisms express proteins in different amounts as conditions change. This regulation often helps the organism adapt: for example, in the presence of a new source of food or energy, the organism may express more of the enzymes for consuming that resource. This study shows that, for bacteria growing in artificial laboratory environments, there is little correlation between when genes are required for optimal growth and when those genes are more highly expressed. Furthermore, this misregulation often reduces the bacteria's rate of growth to a measurable extent. Such drastic misregulation is unlikely to be prevalent in nature, because bacteria can evolve different regulation quite quickly. To explain this discrepancy, we propose that many genes are regulated by signals that are not directly related to their function. This "indirect" control is often not adaptive in artificial environments, which will confuse such a regulatory system, but it probably benefits the bacterium in natural conditions.

Introduction

In bacteria, gene regulation is traditionally thought of as an adaptive or homeostatic mechanism that allows the cell to respond to changing metabolic conditions or to environmental stresses (e.g., [1, 2]).

The underlying rationale is that proteins "should" be made only when needed so as to conserve cellular resources or because the protein's activity is detrimental in other conditions. The classic example is the induction of genes for utilizing lactose in *Escherichia coli* when the carbon source shifts from glucose to lactose.

Although many specific examples of gene regulation appear to be adaptive, it is not clear whether the regulation of the majority of genes is adaptive. We recently compared gene expression to mutant fitness, as measured by assaying the relative growth of thousands of pooled mutants [3]. In the metal-reducing bacterium *Shewanella oneidensis* MR-1, across four conditions, we found little correlation between mutant fitness and gene expression [4]. In other words, most genes were not up-regulated when they were needed for optimal growth. Studies in the model eukaryote *Saccharomyces cerevisiae* also found little correlation between expression and fitness [5–7].

There have been several proposals for why genes might be expressed when they are not needed for fitness or why they might not be induced when they are needed. First, some genes might be expressed in "standby mode" because they will help the bacterium survive if conditions change [8]. Nevertheless, genes that are on standby should still be expressed more highly when they are actually needed, so standby expression should dampen the dynamic range of expression without changing the pattern. Second, proteins that are only needed in modest amounts might be expressed constitutively because the cost of adaptive control, such as the cost of making transcription factors, might exceed the benefit of making less of the protein when it is not needed [9]. Third, horizontally transferred genes, which are common in bacteria, might lack regulation because of insufficient time to evolve appropriate regulation in their current host [10]. On the other hand, many horizontally transferred genes are under complex control by multiple transcription factors [11], regulation can evolve quickly [12,13], and regulation can be conserved across transfer events [11]. Thus, we doubt that horizontal gene transfer could explain why there is little correlation between expression and fitness. Fourth, microorganisms might use one environmental signal to "anticipate" another [14, 15]. For example, for a gut bacterium like *Escherichia coli*, a rise in temperature might indicate that it has been ingested and will soon reach an anaerobic environment [14], so genes for anaerobic respiration might be induced even though they are not immediately useful. It is not clear whether anticipatory control of expression is widespread in bacteria. Finally, the regulation of some genes might be suboptimal or maladaptive because the expression patterns of those genes are not under strong selection. (More precisely, if altered regulation improves relative growth by less than $1/N_e$ per generation, where N_e is the effective size of the bacterial population and the effect on growth is averaged across natural environments, then this altered regulation is unlikely to take over the population.) Selectively-neutral evolution could also account for some of the complexity of gene regulation [16]. However, both regulatory sites [17,18] and the coexpression of genes [19] are usually conserved between closely-related bacteria, which implies that the regulation of most genes is under selection.

To try and understand in more depth whether bacterial gene regulation is adaptive for most genes, we collected genome-wide mutant fitness data and gene expression data from the metal-reducing bacterium *Shewanella oneidensis* MR-1 across 15 matching conditions. We also examined large compendia of (unmatched) fitness and expression data for this bacterium. Based on our findings, we argue that the regulation of most *S. oneidensis* MR-1 genes is maladaptive in the laboratory and suboptimal in the wild:

- Genes that are detrimental to fitness are often highly-expressed instead of being down-regulated when not needed.
- The correlation between relative expression and mutant fitness is weak, as in our previous study [4].
- Many genes are expressed constitutively instead of being controlled by transcription factors, or are regulated by growth rate.
- Even if we exclude constitutive or growth-regulated genes and consider only genes that are important for growth in a subset of conditions, we find only a slight tendency for genes to be more highly expressed when they are important for fitness.

- The genes that have a strong correlation between expression and fitness in our conditions are primarily involved in amino-acid synthesis or carbon-source catabolism – the regulation of most other types of genes does not appear to be adaptive.
- Pairs of genes with closely-related functions, as determined by their annotations and by mutant fitness data, often have divergent expression patterns if they are not in the same operon.
- We found little evidence of anticipatory control.

To test the generality of our findings, we examined the expression and mutant fitness of biosynthetic genes in four diverse bacteria – *S. oneidensis* MR-1, *E. coli* K-12, the ethanol-producing bacterium *Zymomonas mobilis* ZM4, and the sulfate-reducing bacterium *Desulfovibrio alaskensis* G20. In many cases, biosynthetic genes that were required for fitness in minimal media were not upregulated in minimal media. We also collected matched fitness and expression data for *Z. mobilis* ZM4 across 18 conditions, and found little correlation between expression and fitness in *Z. mobilis* ZM4. To explain why there is little correlation between expression and fitness and why much of gene regulation is maladaptive in the laboratory, we propose a model of indirect and suboptimal control.

Results

Matched gene expression and mutant fitness data for S. oneidensis MR-1

We collected genome-wide data on mutant fitness and mRNA abundance for *S. oneidensis* MR-1 grown in 15 matched conditions: Luria-Bertani medium (LB), a defined minimal medium with one of eight different carbon sources added, minimal lactate medium with four different inhibitory compounds added, or anaerobic respiration of fumarate with two different electron donors. For each condition, we measured gene expression from wild-type cells in exponential phase and we measured fitness using two pools of mutants that grew for 6-8 generations (Figure 1A). We obtained both expression and fitness data for 3,247 of the 4,467 protein-coding genes in the genome. (Of the genes that we do not have data for, about half are essential for growth in LB, are under 300 nucleotides, or are repetitive elements such as transposases.) The mutant fitness for each gene is calculated from the $\log_2$ abundance of strain(s) with transposons inserted within that gene, as estimated using a microarray. The fitness values are normalized so that wild-type would have a fitness value of about zero: negative fitness indicates that the mutant strain is sick (relative to wild-type) and that the gene's activity is important for growth in that condition, while positive fitness indicates that the mutant strain has an advantage and that the gene's activity is detrimental in that condition. We will first discuss the relationship between genes' fitness and their absolute level of expression in a given condition, and then examine relative expression across conditions.

Genes important for fitness tend to be highly expressed

As shown in Figure 1B and 1C, during the aerobic growth of *S. oneidensis* MR-1 in a defined medium with lactate as the carbon source, genes that have strong phenotypes when mutated tend to be more highly expressed. Of the genes that are important for fitness (fitness under -0.75), 74% are expressed more highly than the median gene. On average, these important genes are expressed about two-fold higher than other genes (difference in average $\log_2$ levels = 0.85, $P < 10^{-15}$, t test). In 14 of 15 conditions, most of the genes that are important for fitness are highly expressed (Figure 1D). The exception is aerobic growth with acetate as the carbon source, during which genes that are predicted to be involved in the biosynthesis of amino acids, nucleotides, or cofactors [20] are often weakly expressed despite being important for fitness (Figure 1E). This probably reflects the low growth rate of *S. oneidensis* MR-1 on acetate as a carbon source, which implies a low rate of synthesis of cellular components.

Genes are often detrimental to fitness

Many genes are detrimental to fitness during aerobic growth on lactate, i.e., strains with insertions in these genes grow better than most other strains (Figure 1B). Furthermore, detrimental genes tend to be well-expressed (Figure 1B & 1C). Similarly, in all 14 conditions with detrimental genes, the majority of detrimental genes are expressed above the median gene. (There are no strongly-detrimental genes in LB.) The high expression of detrimental genes confirms the fitness data, because it is easier for a gene to exert a strong detrimental effect if it is highly expressed. On the other hand, one wonders why these genes are not down-regulated to reduce their detrimental activity.

To examine this issue more broadly, we identified genes that were detrimental to fitness in a compendium of 195 fitness experiments for *S. oneidensis* MR-1 [4]. Because we were interested in genes that were detrimental to growth or survival, we removed eight experiments that measured motility, leaving us with 187 experiments. To increase sensitivity, we grouped together fitness experiments that had similar patterns (pairwise correlation above 0.75), giving 38 groups. Within each group and for each gene, we required an average fitness above 0.4 as well as statistical significance from combining z scores ($P < 0.01$ after Bonferonni correction for the number of groups). We identified 798 genes (24% of the genes for wich we have fitness data) as significantly detrimental in at least one group of experiments. To validate this result, we examined adjacent pairs of genes that are cotranscribed in the same operon. Genes in the same operon often, but not always, have related functions [21–23], so if one of them is detrimental to fitness then the other should more often be detrimental. Indeed, one gene in an operon pair was much more likely to be detrimental to fitness if the other one was (53% versus 17%, $P < 10^{-15}$, Fisher exact test). This confirms that most of these genes are genuinely detrimental to fitness in our laboratory experiments.

These 798 genes that are detrimental to fitness are not simply selfish genes. Just 18 of them are annotated as potentially selfish elements such as transposases, prophages, or restriction systems. 421 of the detrimental genes (53%) are important for growth or survival in another group of experiments in our compendium (fitness under -0.4 and $P < 0.01$ after Bonferonni correction). Some of the detrimental genes are involved in motility, which is consistent with previous reports [4, 24, 25], but we doubt that motility can account for most of the detrimental genes. We previously measured mutant motility in *S. oneidensis* MR-1 by assaying the abundance of mutant strains that reached the outer ring of a soft agar plate [4]. (These are the same experiments that were removed from the fitness compendium because they did not measure growth or survival.) 34% of the 798 detrimental genes have a motility "fitness" of under -0.4, as compared to 13% of other genes. Although the detrimental genes are enriched in motility genes ($P < 10^{-15}$, Fisher exact test), motility and selfishness together only explain around a third of the detrimental genes. In any case, the regulation of the 421 genes that are sometimes detrimental and sometimes important for growth – 13% of the genes that we have fitness data for – is suboptimal, at least in our laboratory conditions, as these genes "should" be repressed when they are detrimental to growth.

Relative expression is little correlated with fitness

We then asked if genes are upregulated when they are needed. One way to test this is to compare differential expression and the difference of mutant fitness between two conditions. We previously reported three such comparisons and found little correlation [4]. Here, we repeat this test for 14 comparisons derived from our 15 conditions, with aerobic growth in minimal lactate media as the common control. For example, Figure 2A shows a comparison of differential expression and fitness for aerobic growth in acetate versus lactate. If there was a strong relationship between expression and fitness, then genes that are more important for fitness on acetate than on lactate (i.e., a fitness difference below zero) would also be upregulated (i.e., an expression $\log_2$ ratio above zero). Overall, there would be a strong negative correlation between differential fitness and relative expression. Instead, the correlation is statistically significant but is very weak ($r = -0.15$, $P < 10^{-15}$; Figure 2A). As mentioned above, biosynthetic genes might be needed at lower levels on acetate than on lactate, while still being important for fitness in both

conditions; nevertheless, after removing biosynthetic genes, the correlation is still very weak (r = -0.08). In all of the comparisons, the correlation between differential expresion and fitness is weak ($r = -0.15$ to $+0.11$).

In most of these comparisons, genes that are important for fitness in just one of the two conditions do tend to change expression in the expected direction (e.g., Figure 2B). However, over a third of these differentially-fit genes change expression the "wrong" way (i.e., lower expression on acetate for genes that are important only on acetate or lower expression on lactate for genes that are important only on lactate). The two distributions of expression changes (for the two types of differentially-fit genes) overlap considerably, which can be quantified with the Kolmogorov D statistic, which depends only on the relative ranks of the values and ranges from 0 for identical distributions to 1 for distributions that do not overlap. For acetate versus lactate, $D = 0.23$. For all of our comparisons, there is much overlap in the distributions of relative expression between genes that are sick only in one condition or only in the other (Figure 2C, $D = 0.12$ to 0.71). In a few conditions, genes that are differentially important for fitness are just as likely to change expression in the "wrong" direction. For example, of 22 genes that are important for fitness with copper stress but not without it, 12 are downregulated on copper stress (Figure 2D).

Conversely, many of the genes with large changes in expression are not important for fitness. In the comparison of acetate and lactate, of 114 genes that changed expression by four-fold or more, 70 (61%) have little effect on fitness in either condition (both fitness values between -0.4 and 0.4). In most of the other comparisons, this proportion was even higher (up to 87% in acid stress). Because it is difficult to measure small differences in fitness, it is possible that the change in expression of these genes is adaptive because these genes have subtle effects on fitness. However, given the weak overall correlation between differential expression and fitness, this seems unlikely. Genes with large changes in expression could also lack phenotypes because of genetic redundancy, but we expect that this is not a major factor because paralogs in *S. oneidensis* MR-1 often have phenotypes when mutated [4].

Another way to ask if genes are upregulated when they are needed is to look at the correlation, for any given gene, between expression level and mutant fitness across the 15 conditions (see examples in Figure 3A). If a gene is more highly expressed when it is important for fitness, then we should see a strong negative correlation (e.g., *tyrA* in Figure 3A). Instead, the distribution of fitness-expression correlations for all genes is about the same as if we shuffle the data and compare a gene's fitness pattern to another random gene's expression pattern (Figure 3B). The actual distribution is significantly different from the shuffled distribution (P = 0.01, Kolmogorov-Smirnov test) but the difference is slight, with average correlations of 0.00 and 0.01, respectively. Overall, when we compare differential expression and fitness, either by selecting pairs of conditions or by examining each gene across all 15 matched conditions, we find that they are weakly correlated. This suggests that the regulation of most genes is not adaptive under our laboratory conditions.

Genes with close functional relationships are often not coregulated

To confirm that gene expression patterns are often not correlated with a gene's function, we examined the coexpression of genes that have closely-related functions but are not in the same operon. Using a compendium of 195 fitness experiments for *S. oneidensis* MR-1 [4], we identified 240 pairs of genes that were highly cofit (correlation of fitness above 0.8), were annotated with the same TIGR subrole [20], did not belong to the same predicted operon [26,27], and were not nearby each other in the genome (not within 10 genes of each other). When we examined the coexpression of these functionally-related pairs across 329 expression experiments for *S. oneidensis* MR-1, we found that they have only a moderate tendency to be coexpressed (Figure 3C). For example, 83% of the operon pairs have a coexpression of 0.5 or higher, but just 43% of the functionally-related pairs do. Furthermore, according to gene regulation that was predicted via comparative genomics and manually compiled in RegPrecise [28], these functionally-related pairs are usually not coregulated: of the 240 pairs, there is a regulatory prediction for at least one gene

among 97 pairs, and both genes are predicted to be regulated by the same transcription factor in only 7 cases.

To test this more carefully, we manually examined the 76 pairs of genes with a close functional relationship but little coexpression ($r < 0.3$). Thirty six of the 76 pairs had known functional differences or showed differences in fitness in a few conditions that might explain their limited coexpression (Table S1). For example, genes for both proline and arginine synthesis have the TIGR subrole "Amino acid biosynthesis: Glutamate family" and show similar fitness in most, but not all, of our conditions. It is not surprising that they might be regulated differently. These pairs reflect the limited resolution of the functional classification. Another 18 pairs of genes were from flagellar operons *fliKLMNOPQR-flhB*, *flgL1-flaG-fliD-SO_3234-fliS*, *flgFGHIJ-SO_3239.3-SO_3239.2-flgL2-flgL3*, *flgBCDE*, and *flgAMN*. Some of the differences in expression of these genes might reflect the sequential activation of different stages of assembly of the polar flagellum, which has been studied in detail in related bacteria (e.g., [29, 30]). However, in *Pseudomonas aeruginosa*, seven of these 18 pairs of genes are co-regulated and are in the same "class" of transcripts [30], so it is not clear that these genes are needed at different times. The remaining 22 pairs of genes were from operons with closely-related functions for which there was no apparent reason for the expression to differ. Specifically, these pairs of genes were from aromatic amino acid synthesis operons *aroA*, *aroC*, *aroE*, *aroQ*, and *aroKB*; menaquinone synthesis operons *menA*, *menB*, *menF*, and *menDHCE*; branched-chain amino acid synthesis operons *ilvGMDA*, *ilvC*, and *ilvE*; pyrimidine synthesis genes *pyrC*, *pyrD*, and *pyrF*; methionine synthesis operons *metBL* and *metC*; lipid A synthesis genes *lpxL* and *lpxM*; mismatch repair genes *mutL* and *mutS*; and chromosome separation genes *xerC* and *xerD*. Among these genes, *mutL*, *pyrD*, *pyrF*, and *xerC* are in operons with functionally-unrelated genes, while the other genes listed individually are transcribed separately, as determined using high-resolution "tiling" microarrays and 5'-end RNA sequencing (see Materials and Methods). If gene regulation evolves to an optimum, then it is difficult to explain why the regulation of these genes or operons would be different, especially for genes that are not cotranscribed with functionally-unrelated genes. Thus, the lack of coexpression for these genes with closely-related functions appears to be suboptimal.

Suboptimal control via constitutive or growth-rate regulation of many genes

One explanation for why there is little correlation between fitness and expression is that some genes are expressed constitutively and are not under adaptive regulation. Indeed, in the compendium of 329 expression experiments for *S. oneidensis* MR-1, many genes show little change in their expression. We chose to classify genes with a standard deviation of their $\log_2$ expression ratio of under 0.5 as constitutive; this accounts for about one sixth of the genes in the genome. Although our threshold is somewhat arbitrary, these constitutive genes are much less likely than other genes to have predicted regulation in RegPrecise [28] (3.6% versus 16.1%, $P < 10^{-15}$, Fisher exact test). This supports the idea that most of these constitutively-expressed genes are not regulated by specific transcription factors and are not subject to adaptive control. (RegPrecise does not attempt to predict binding sites for the major sigma factor (σ^{70}) or for nucleoid proteins.) The constitutive genes are functionally diverse, and most types of genes are represented. The only TIGR subrole that is significantly depleted is electron transport (false discovery rate under 0.05, Fisher exact test).

We also hypothesized that many genes would be regulated by growth rate, because at higher growth rates, a higher proportion of cellular resources are devoted to transcription and translation [31]. We examined the expression patterns of 24 essential protein components of the ribosome (*rplBCDFJLM-NORTWX* and *rpsBEHIJLMNPQS*) and as expected, these genes are quite coexpressed, with a median pairwise correlation of 0.83. We used the average expression profile of these ribosomal genes to identify other putatively growth-correlated genes. Specifically, we identified 391 genes whose coexpression with the profile is above 0.5, including all of the original 24 genes. These "growth-regulated" genes are only slightly less likely than other genes to be regulated by specific transcription factors according to RegPrecise (10.7% vs. 14.4%, $P = 0.054$, Fisher exact test). Nevertheless, we can confirm that they are

growth-regulated by examining their promoter sequences. In *E. coli* and presumably in *S. oneidensis* MR-1 as well, the growth regulation of ribosomal protein genes is mediated by the alarmone ppGpp and the DksA protein as part of the stringent response [32]. DksA binds to RNA polymerase and alters the efficiency of transcription initiation depending on various factors including the concentration of the first (initiating) nucleotide and a GC-rich "discriminator" between the -10 box and the initiation site [33–35]. We used a combination of high-resolution "tiling" microarrays and 5' RNA-Seq to map the exact 5' ends of transcripts for 1,236 genes or operons from *S. oneidensis* MR-1. We found a substantial difference in the initiating nucleotides between growth-regulated and other transcripts: just 25% of growth-regulated transcripts begin with adenosine, while 51% of other transcripts do ($P < 10^{-7}$, Fisher exact test). The putative growth-regulated promoters also have a higher GC content at positions -4 to -1 than other promoters do (68% vs. 55%, $P < 10^{-5}$, t test). Thus, many of the putative growth-regulated promoters in *S. oneidensis* MR-1 are affected by the stringent response. As with the constitutive genes, the growth-regulated genes are functionally diverse: although they are enriched for translation-related genes, the only TIGR subrole that is significantly depleted is anion transport (false discovery rate under 0.05).

Not surprisingly, constitutive genes and growth-regulated genes do not show a correlation between fitness and expression: across our 15 matched conditions, the two groups have mean fitness-expression correlations of 0.01 and 0.00, respectively (both $P > 0.5$, t test). Together these account for 21% of the genes for which we have both fitness and expression data, so constitutive or growth-regulated expression could explain the lack of adaptive control for many genes. To try to identify a subgroup of genes that might show more correlation between fitness and expression, we considered only genes that strongly affect fitness in at least one of our 15 matched experiments (maximum $|\text{fitness}| > 0.75$). As shown in Figure 3D, among genes that affect fitness, constitutive and growth-correlated genes still show no fitness-expression correlation (both $P > 0.4$, t test), but some of the other genes do (mean -0.11, $P < 10^{-13}$, t test). Of the other genes that affect fitness, 16% have strong negative fitness-expression correlations of under -0.5, and many of those are involved in amino acid synthesis (Figure 3E). For example, of the 60 genes with a fitness-expression correlation under -0.5 and an annotated TIGR subrole, 31 (52%) were involved in amino acid biosynthesis. No other functional category was enriched in genes with strong fitness-expression correlations, but 11 of these genes are involved in the catabolism of the carbon sources we used (fadAB, deoC, gnd, edd, zwf, astB, nagABK, and SO_3774). The traditional view that much of gene regulation is adaptive could be an over-generalization from the extensive study of gene regulation of biosynthetic and catabolic pathways.

Little evidence for anticipatory control

Another possible explanation for the weak correlation between expression and fitness is that the bacterium is anticipating growth in a different environment [14, 15]. We systematically looked for evidence of anticipatory control by considering all pairs of our conditions. Given conditions A and B, if the organism uses A to anticipate B, then genes that are required for growth on B but not on A should be upregulated on A (relative to a control condition) as compared to genes that are not required for growth in either condition. We used the median expression across the 15 conditions as the control and tested the 203 pairs of conditions that have at least 10 differentially-fit genes. We found only two cases of potential anticipation that were statistically significant ($P < 0.01$, Wilcoxon test with Bonferonni correction).

The most significant effect was that growth on CAS, a mixture of amino acids, "anticipated" growth on gelatin (corrected $P < 10^{-8}$). Rather than being a form of anticipatory control, we suspect that *S. oneidensis* MR-1 cannot distinguish growth on the peptides in gelatin from growth on amino acids, so it expresses genes for taking up peptides whenever amino acids are present. Of the 15 genes that were sick on gelatin but not on CAS and that were up-regulated two-fold or more on CAS, three are involved in peptide uptake (SO_1822, SO_3194.1, and SO_3195).

The other significant effect was that aerobic growth on pyruvate anticipated anaerobic growth on N-acetylglucosamine (NAG) with fumarate as the electron acceptor (corrected $P < 10^{-6}$). Of 33 genes

that are important for fitness with NAG/fumarate but not on pyruvate, 7 genes were up-regulated by 1.5-fold or more on pyruvate. Three of these genes form a hydrogenase operon (SO_2099:SO_2097) that is predicted to be regulated by Crp and Fnr [28], and three of the other four genes are predicted to be regulated by Crp or Fnr (ccmC, ccmA, and ccmH). Crp and Fnr are both regulators of anaerobic respiration in this organism [36, 37], and both the Crp and Fnr regulons are upregulated on pyruvate (both $P < 10^{-8}$, t test) so we speculate that oxygen levels might drop during batch aerobic growth on pyruvate. Alternatively, there may be another signal for up-regulating these genes. In any case, anticipatory control is not widespread among our 15 conditions and does not seem to explain why we observe little correlation between fitness and mRNA expression.

Variation in expression during the growth phase does not explain the lack of correlation with fitness

Another potential reason why we see little agreement between expression and fitness is that we measured expression at one time during the growth curve (in mid-exponential phase), while our fitness data reflects the importance of the gene throughout the growth curve. For example, if a gene is important for the early adjustment to growth in a new condition but not afterwards, then at the end of the experiment, the mutant strains would have reduced abundance and the gene's fitness would be negative, yet it would be adaptive for the gene to be less-expressed in mid-exponential phase. In a previous study we examined growth curves for 48 *S. oneidensis* MR-1 mutants with a variety of fitness values [4]. Just two mutants grew at a normal rate but with a long lag, and most fitness defects were reflected in the growth rate during mid-exponential phase. Because most genes that affect fitness are important for growth during exponential phase when we collected samples for gene expression, growth phase effects are unlikely to explain why there is little correlation between expression and fitness.

To more directly test how the relationship between expression and fitness might vary with the growth phase, we measured expression at various points in time during batch growth in rich media (LB) or in defined medium with lactate or N-acetylglucosamine (NAG) as the carbon source. The correlation between differential expression and fitness (computed as in Figure 2A) varied across time points, but was never dramatically tighter than in our original experiments. For lactate versus LB, the original correlation was -0.11 and the best correlation during the time course was -0.25; for lactate versus NAG, the original correlation was -0.06 and the best was -0.11; and for NAG versus LB, the original correlation was -0.25 and the best (during the time course) was -0.21. The correlation between differential expression and fitness also remained moderate if we used the maximum expression during each time course (correlations of -0.18 for lactate versus LB, -0.06 for lactate versus NAG, and -0.14 for NAG versus LB, respectively). Thus, the time at which we measured expression does not explain the low correlation between differential expression and fitness.

Repression of biosynthetic pathways in rich media is not the norm

To extend our analysis to diverse bacteria, we compared the expression and fitness of biosynthetic genes between rich and minimal media in four organisms: *Escherichia coli* K-12, *Shewanella oneidensis* MR-1, the ethanol-producing bacterium *Zymomonas mobilis* ZM4, and the anaerobic sulfate-reducing bacterium *Desulfovibrio alaskensis* G20. As shown in Figure 4, auxotrophic genes – genes that are annotated in biosynthetic pathways [20] and are important for fitness in minimal media but not in rich media – tend to be upregulated on minimal media in *E. coli* K-12 and in *S. oneidensis* MR-1 (average $\log_2$ ratio of 1.5 and 0.84, respectively, with $P < 10^{-15}$ and $P < 0.001$, t test). However, in *Z. mobilis* ZM4 or in *D. alaskensis* G20, auxotrophic genes are not upregulated on minimal media (both $P > 0.3$, t test).

Surprisingly, in *S. oneidensis* MR-1, 28 of the auxotrophic genes are down-regulated in minimal media, and 15 of these are involved in nucleotide synthesis. These genes are scattered across 11 different operons – *guaBA*, *purC*, *purEK*, *purF*, *purHD*, *purL*, *purMN*, *pyrC*, *pyrD*, *pyrE*, *pyrF* – so this pattern has evolved

independently many times. *pyrD* and *pyrE* are in operons with functionally-unrelated genes, but there is no obvious reason why the other nine operons are not regulated by nucleotide availability. The expression time courses for LB, lactate, and NAG confirm that the 15 nucleotide synthesis genes are more highly expressed during log phase growth in LB – which contains nucleotides – than at any phase of growth in defined media. Although mutants in guaBA do show a mild growth defect in LB, which suggests that their activity might be required, mutants in the other nucleotide synthesis genes do not. Thus, in *S. oneidensis* MR-1, the expression of nucleotide synthesis genes does not respond to the availability of nucleotides or the cell's requirements for these genes.

We propose that *E. coli* K-12 has evolved direct regulation of biosynthetic pathways by the relevant end products so that it can efficiently utilize many different carbon sources, including amino acids and nucleotides. In particular, the switch between degrading and synthesizing these compounds may require regulation to avoid futile cycles in metabolism. In contrast, *S. oneidensis* MR-1 is adapted for utilizing amino acids but not nucleotides: it does grow on DNA or on a few nucleosides as carbon sources, but poorly, and it cannot utilize nucleobases [38, 39]. A genome-scale metabolic model suggests that during growth on adenosine or inosine, it degrades the ribose or deoxyribose portion and secretes the nucleobases [40]. If *S. oneidensis* MR-1 is not adapted to utilizing nucleobases, this might explain why it does not control the expression of these synthesis pathways by nucleotide availability. Finally, *Z. mobilis* ZM4 and *D. alaskensis* G20 do not, as far as we know, use amino acids or nucleotides as carbon sources and may not have encountered high levels of amino acids or nucleotides often enough to evolve transcriptional regulation of these pathways in response to those compounds. Overall, we found that biosynthetic pathways are often not downregulated when their end products are available.

No correlation between expression and fitness in Zymomonas mobilis ZM4

Finally, to test the relationship between expression and fitness in another bacterium in diverse conditions, we collected mutant fitness data and gene expression data for *Zymomonas mobilis* ZM4 across 18 conditions. As *Z. mobilis* ZM4 can only use a few sugars as carbon sources, we studied growth in rich and minimal media and in various stresses. As in *S. oneidensis* MR-1, we found little correlation between expression and fitness in *Z. mobilis* ZM4. Unlike in *S. oneidensis* MR-1, in *Z. mobilis* ZM4 there was no difference between the distribution of per-gene fitness-expression correlations and the shuffled distribution ($P > 0.5$, Kolmogorov-Smirnov test with 1,568 genes and 1,568 controls). The mean correlations were 0.007 and 0.006, respectively. After removing genes without fitness effects, constitutively-expressed genes, and growth-regulated genes, the mean correlation remained at 0.007. Our inability to find any correlation between expression and fitness in *Z. mobilis* ZM4 could reflect the rather artificial conditions we used, less careful matching of the experimental conditions for the two assays than in *S. oneidensis* MR-1, or a simpler regulatory system – *Z. mobilis* ZM4 has just 65 transcription factors while *S. oneidensis* MR-1 has 243.

Discussion

We have shown that in diverse bacteria, there is little correlation between when genes are important for fitness and when they are more highly expressed. We also tested two potential explanations for why the correlation is low: (1) a mismatch between when we measured expression and when we measured fitness, or (2) that unnecessary proteins are upregulated in one condition in anticipation of another condition. We did not find support for either of these explanations. Although we cannot quantify what proportion of genes lack adaptive regulation, our results do not seem consistent with the traditional view that most of bacterial gene regulation is adaptive.

Transcriptional control and protein levels

A major caveat in our study is that we have analyzed mRNA levels rather than protein levels. It is not clear to what extent the variation in a protein's concentration is determined by the variation in its mRNA's concentration. However, we did show that roughly a quarter of all genes in *S. oneidensis* MR-1 are detrimental to fitness in at least one condition, and that detrimental genes tend to be highly expressed. This proves that the RNA expression of these genes is suboptimal in the laboratory.

A model of indirect control

We propose that in bacteria, most genes are under "indirect" control – their regulator(s) sense signals that are not directly related to the gene's function, but in the environment, there is a correlation between those signal(s) and whether the gene's activity is beneficial. For example, for bacteria that alternate between gut and soil environments, low temperature might correlate with low levels of nutrients, but these correlations will not be found in the laboratory. We suspect that many genes are under such indirect control because a functionally-relevant regulator is not available: for example, *S. oneidensis* MR-1 has 4,467 protein-coding genes and around 2,800 transcripts but only 243 transcription factors (5.4% of proteins). (For comparison, in the typical bacterium, 4.2% of proteins are predicted to be transcription factors [41].) Because there are ten times more operons than transcription factors, for most operons there are no transcription factors in the genome that sense functionally-relevant signals. Thus, direct and optimal control of most genes does not seem feasible.

Many aspects of bacterial regulation are consistent with indirect control. First, we showed that many genes, with diverse functions, are expressed constitutively or are regulated by growth rate. Such simple control is easy to evolve, as no specific transcription factors are required. Second, many genes are regulated by a handful of highly-expressed "global" regulators that regulate diverse and sometimes functionally-unrelated genes [42]. Because global regulators are present at high concentrations, they will bind at low-affinity sites that require relatively-little information to specify [43, 44], so these sites should evolve more readily than binding sites for other regulators [12, 13]. This might explain why binding sites for global regulators have evolved upstream of functionally-diverse genes. Third, theoretical analysis of the transcriptional regulation of biosynthetic pathways suggests that the optimal design is for them to be regulated by their end product, but many pathways are instead regulated by transcription factors that sense metabolic intermediates [45]. This is consistent with our view that sensors for the optimal signals might not be available.

Indirect control may explain why the regulation of most genes seems suboptimal in the laboratory – correlations that exist in the environment will probably not be maintained in our experiments. Intuitively, we are confusing the bacteria by growing them in unfamiliar conditions such as high nutrient levels, high cell densities, pure carbon sources, no competition from other microorganisms, and no predation. For example, consider the 13% of genes in *S. oneidensis* MR-1 that have provably suboptimal control in the laboratory, that is, that are detrimental for fitness (above 0.4) in some conditions but important for fitness (below -0.4) in others. A fitness value of 0.4 implies that improved regulation would have a selective advantage of 3% per generation – a very strong selective pressure. Indeed, laboratory evolution experiments often find beneficial mutations in global regulators [46], which is consistent with the idea that the discrepancies we observe in the laboratory do not occur in the wild (or else these mutations would have evolved already). We also note that fitness costs of 3% per generation are far too high to be explained by the waste of cellular resources in making unneeded protein, as few if any proteins account for 3% of total expression; rather, the activity of these genes reduces growth in some of our laboratory conditions and improves growth in others.

According to our model, gene expression responses will be more adaptive if examined under natural conditions. Measuring gene expression during slow growth at low cell densities in the presence of other microorganisms seems challenging. Nevertheless, given the rapid rate of improvements in DNA and RNA

sequencing, we hope that it will soon become feasible. In the meantime, although we are convinced that control is often indirect, it is difficult to know if control is anticipatory [14,15], simple because the cost of control would not be worth it [9], or truly suboptimal under natural conditions. Given that functionally-related genes often have different expression patterns, and that we found little evidence of anticipatory control, we suspect that control is truly suboptimal in the wild, although the fitness costs of suboptimal control might be much smaller in the wild than in the laboratory.

Additional evidence that bacterial gene regulation might not be optimal in the wild comes from studying operon structures and their evolution. First, many operons contain functionally-unrelated genes [21–23]. In the stomach bacterium *Helicobacter pylori*, operons consist predominantly of functionally unrelated genes [26,47]. It is not clear how the regulation of genes in these operons can be optimal. Second, although operons tend to be conserved across related bacteria [48, 49], operons are rarely conserved between distantly-related bacteria [50]. When operon structures change, gene expression patterns change as well, so it seems unlikely that gene regulation is optimal both before and after the change [23].

Another question is what limits the number of transcription factors in bacterial genomes. There is a roughly linear relationship between the number of proteins encoded by a bacterial genome and the proportion of genes that encode transcription factors [51]. The relatively small number of transcription factors in smaller bacterial genomes suggests that the benefits of additional control would be less than the costs or would be too small for selection to operate. This might reflect the adaptation of bacteria with small genomes to narrow niches. Bacteria with large genomes contain many horizontally acquired genes, and these can be acquired together with transcription factors to regulate them, as the regulator is often nearby in the genome [11]. So, why aren't operons acquired together with regulators more often? One possibility is that there would be interference between transcription factors with similar DNA binding preferences (similar to the theory of [52]). If it is difficult to acquire a transcription factor that senses the relevant signal, it might take a long time to evolve a new one.

Alternative models

Another explanation for the weak correlation between expression and fitness is that many promoters are poorly "insulated" from environmental factors [53]. Even if genes are regulated by transcription factors that sense functionally-relevant signals, their expression also fluctuates due to irrelevant differences in environmental conditions [53]. These fluctuations might be due to non-specific binding of other transcription factors at weak sites that evolve neutrally and are not deleterious enough for selection to remove them [16]. Or the concentration of active transcription factor might fluctuate due to factors besides the signal that the transcription factor senses.

Poor insulation is like indirect control in that the gene's expression responds suboptimally to irrelevant signals, but the effect is proposed to evolve neutrally rather than in response to environmental correlations. Although the two models have many similarities, we showed that constitutive expression and regulation by growth rate are widespread, which does not fit the insulation theory. On the other hand, when we considered genes that have a close functional relationship but are not in the same operon, we saw more coexpression than we might expect from the slight correlation between expression and fitness for most genes (e.g., compare Figure 3C and 3D). This might be explained by poor insulation – if two promoters are responding to transcription factors that sense relevant signals, but the concentrations or activities of those transcriptions factors are affected by irrelevant changes in growth conditions, then expression from those promoters would be well-correlated with each other yet fitness-expression correlations would be modest.

Another possible reason for the weak correlation between expression and fitness is that optimal control requires complex combinatorial regulation. Among genes with characterized regulation in *E. coli* [54], 962 of 1,641 genes (59%) are regulated by more than one transcription factor. One possible reason for why combinatorial control is widespread is to make up for the limited number of sensors. In any case, we speculate that combinatorial logic might perform poorly in laboratory conditions. For example, even

if the sensed signals are functionally relevant, the way in which they are combined might be adapted to natural conditions. We also suspect that combinatorial control implies a rugged fitness landscape for selection on the promoter region, which might make it difficult for optimal control to evolve.

Conclusions

We found that in diverse bacteria, there is little correlation between mutant fitness and gene expression. Most genes are not upregulated when they are needed, and conversely, many genes that are detrimental to growth are highly expressed, which implies that they are not downregulated when they are harmful. For many genes, the lack of correlation seems to reflect constitutive expression or regulation by growth rate. Because bacterial genomes have far more operons than they have transcription factors, we propose that most bacterial genes cannot be regulated by functionally-relevant signal(s), so their regulation is indirect and suboptimal. Besides giving a new perspective to the evolution of gene regulation, our results lead us to doubt whether gene expression data alone will reveal the key genes or pathways that allow bacteria to grow or survive in diverse conditions.

Materials and Methods

Fitness and expression data for S. oneidensis MR-1

We collected matched mutant fitness and gene expression data for *S. oneidensis* MR-1 (ATCC 700550) in 15 conditions: aerobic growth in Luria-Bertani broth; aerobic growth in defined minimal media with 8 different carbon sources (20 mM D,L-lactate, 20 mM pyruvate, 10 mM acetate, 20 mM N-acetylglucosamine (NAG), 5 mg/mL mixed amino acids (CAS), 1 mg/mL gelatin, 0.5% Tween-20, or 7.5 mM inosine); aerobic growth in defined lactate medium with four different stresses (70 μM copper(II) chloride; 1 mM sodium nitrite; 1.5 μM nalidixic acid, an inhibitor of DNA gyrase; or acid stress at pH 6); and anaerobic growth in a defined medium with 20 mM D,L-lactate or 20 mM NAG as the carbon source and 30 mM fumarate as the electron acceptor. Our defined medium contained 30 mM PIPES buffer, salts (1.5 g/L NH_4Cl, 0.1 g/L KCl, 1.75 g/L NaCl, 0.61 g/L $MgCl_2{\cdot}6H_20$, 0.6 g/L NaH_2PO_4), Wolfe's vitamins, and Wolfe's minerals, at pH 7. For the stress experiments, the carbon source was 20 mM D,L-lactate. For growth at pH 6, we used 30 mM MES buffer instead of PIPES. All *S. oneidensis* MR-1 samples were grown at 30°C with shaking at 200 rpm.

For each condition, we collected gene expression data from wild-type cells and fitness data from two pools of transposon mutants, and all three cultures for a given condition were initiated at the same time with the same media. Samples for gene expression were collected in exponential phase, and samples for fitness were collected after 6-8 doublings of the population. In pilot experiments, it made little difference whether we collected fitness data in late exponential phase or in stationary phase (data not shown).

For three conditions, we also measured gene expression during batch growth. We collected cells at varying times after inoculation of batch aerobic growth at OD_{600} of 0.1 on minimal lactate medium (7 samples and maximum OD=0.55), minimal NAG medium (6 samples and maximum OD=1.6), and LB (7 samples and maximum OD=4.0).

For fitness experiments, strain abundance was quantified using a microarray as described previously [4]. Briefly, we extracted genomic DNA, used PCR to amplify the tags that "barcode" each strain, hybridized the amplified tags to a Affymetrix 16K TAG4 array, and scanned the array [55]. Each strain's barcode actually contains two different tags – we amplified the "uptags" from one pool and the "downtags" from the other pool, mixed them together, and hybridized them to one array.

As shown in Figure 1A, fitness values for each strain were computed from the $\log_2$ ratio of abundance after growth versus the start of the experiment. Fitness values for each gene were the average of the per-strain values. Because we use two pools of mutants that are grown and assayed separately, and because

some strains are present in both pools, we can verify the reliability of a fitness experiment by asking whether strains gave similar values from both pools. We quantified this by looking at the correlation of these strains' fitness values across the two pools. In our typical fitness experiment for *S. oneidensis* MR-1, the correlation of strain fitness values was 0.92, and all experiments had correlations above 0.8 except for NAG/fumarate (r = 0.66). In the NAG/fumarate experiment, pairs of genes in the same operon did have well-correlated fitness values ($r = 0.66$, as compared to $r = 0.63$ in our typical experiment).

To quantify gene expression, we used a 12-plex Nimblegen microarray in which each sector has 122,643 spots and 40,881 distinct probes as described previously [4]. Briefly, we used RNAProtect (Qiagen), isolated total RNA (RNAeasy mini kit, Qiagen), prepared first-strand labeled cDNA (SuperScript Plus Indirect cDNA Labeling Module, Invitrogen), and hybridized the labeled cDNA to the microarray according to Nimbelegen's instructions. Within each experiment, the log-level of expression of genes in the same operon was highly correlated (r = 0.75-0.88 for matched experiments, but growth curve experiments had values as low as 0.62).

Compendium of expression data for S. oneidensis MR-1

We obtained 371 expression experiments from the MicrobesOnline web site [27], derived primarily from [4,56,57] and similar works. We removed experiments and genes with a high proportion of missing values, leaving data for 3,844 genes across 329 experiments.

Transcript structures of S. oneidensis MR-1

We grew *S. oneidensis* MR-1 in minimal lactate media and collected high-resolution "tiling" microarray data and performed RNA sequencing targeting the 5' ends of transcripts, using protocols described previously [58]. Briefly, we extracted RNA from frozen cell pellets using RNeasy miniprep columns with DNase treatment (Qiagen), confirmed RNA quality with Agilent bioanalyzer, and depleted ribosomal RNA with MICROBExpress kit (Ambion). For the tiling experiment, we then created labeled first-strand cDNA with SuperScript (Invitrogen) to hybridize to an a microarray (Nimblegen) with 2.01 million probes of 60 nucleotides each. For the 5' RNASeq experiment, we used terminator 5'-phosphate-dependent exonuclease (Epicentre) to remove degraded transcripts, converted 5'-triphosphate to 5'-monophosphate ends with RNA 5' polyphosphatase (Epicentre), ligated adapters onto the 5' end with T4 RNA ligase (Ambion), used random hexamer primers that also included a sequencing adaptor to create cDNA, and used PCR amplification to enrich for DNA that contained both adaptors (see [58] for details). The 5' RNA-Seq data (Illumina) gave 21.5 million reads that mapped uniquely to the genome. To identify transcript starts, we combined local peaks in the 5' RNA-Seq data with sharp rises in the tiling data [58]. Specifically, we used local peaks in the 5' RNA-Seq data that had at least 50 reads and we required these starts to be within 30 nucleotides of a sharp rise in the tiling data that had a local correlation to a step function [59] of at least 0.8. We associated a transcript start with a gene if it was up to 200 nucleotides upstream of the 5' end of the gene. For transcript start analyses, we considered only genes on the main chromosome.

Fitness and expression data for Z. mobilis ZM4

Our standard growth media for *Z. mobilis* ZM4 (ATCC 31821) was aerobic growth at 30°C in a rich medium with 25 g/L glucose, 10 g/L yeast extract, and 2 g/L KH_2PO_4. We collected fitness and expression data for *Z. mobilis* ZM4 grown in this condition and with various inhibitory compounds added, namely 0.45% furfuryl alcohol, 4 mM 4-hydroxybenzaldehye, 5-10 mM 3-hydroxybenzoic acid, 7% ethanol, 0.09%-0.12% acetic acid, 0.2% acetic acid, 7.5 mM 5-hydroxymethylfurfural, 1% butanol, 9.9-12.5 mM furoic acid, 17-26 mM levulinic acid, 0.1-0.2 M NaCl, 0.0004-0.00055% hydrogen peroxide, 2.5 mM vanillin, or a complex stress provided by 6-8% hydrolyzed plant material. Some of the concentrations are

given as ranges because the fitness experiments were done at more than one concentration or at a different concentration from the expression experiments. If the fitness experiments were done at more than one concentration or more than once then we averaged them. The correlation of the per-gene fitness values from experiments with different concentrations of the same inhibitor was above 0.8 in 23 of 24 cases. We also used a defined medium containing 20 g/L glucose, salts, and vitamins [60]. Fitness was measured using a similar approach as in *S. oneidensis* MR-1; the two pools of transposon insertions that we used will be described in more detail elsewhere (J.M.S. *et al.*, in preparation). In the typical experiment for *Z. mobilis* ZM4, the correlation of strain fitness values between the two pools was 0.94, and all experiments had correlations above 0.8. We measured gene expression with a microarray from Nimblegen with 51,851 probes for 1,882 genes, using the same protocols as for *S. oneidensis* MR-1. Within each experiment, the log-level of expression of genes in the same operon was well-correlated (r = 0.78-0.88).

Fitness and expression data for D. alaskensis G20

We grew *D. alaskensis* G20 (provided by Terry Hazen, University of Tennessee, Knoxville) anaerobically at 30°C in a defined lactate-sulfate medium (LS4D) and in a similar medium supplemented with yeast extract (LS4), as described previously for *D. vulgaris* Hildenborough [58]. We collected fitness data using a similar approach as in *S. oneidensis* MR-1; the two pools of transposon insertions that we used will be described in more detail elsewhere (J.V.K. *et al.*, in preparation). Unlike in *S. oneidensis* MR-1 or *Z. mobilis* ZM4, we used separate chips to assay the two pools for a given condition: for each sample, we amplified both the uptags and the downtags and we hybridized those to the same array. We averaged the $\log_2$ intensities of the up- and down-tags together before further processing. In both rich and minimal media, strain fitness was highly consistent between the two pools ($r = 0.94$ and $r = 0.97$, respectively).

We measured gene expression in *D. alaskensis* G20 with a high-resolution "tiling" microarray (Nimblegen) with 2.1 million 60-mer probes, using the same protocols as with the *S. oneidensis* MR-1 tiling array. We considered only probes for the coding strand of genes, we used quantile normalization to put the two data sets into the same distribution, and we averaged the normalized $\log_2$ intensities across the probes for each gene. Genes in the same operon had highly-correlated expression differences between rich and minimal medium ($r = 0.87$).

Analysis of mutant fitness data

In previous work on fitness data from *S. oneidensis* MR-1 [4], we normalized the fitness values so that the median strain had a fitness of zero. Because there can be differential efficiency in extracting DNA of different sizes, we did this separately for the main chromosome and the megaplasmid. We had found that some experiments had significant effects depending on which microplate the strain was grown on during assembly of the pools, so we also normalized the data so that each "pool plate" had a median fitness of zero. Here, we used pool-plate normalization for *S. oneidensis* MR-1 and for *Z. mobilis* ZM4, but it was not needed for *D. alaskensis* G20.

We also identified a small trend by chromosome position in some fitness experiments. Specifically, there was sometimes a correlation between fitness and the distance from the origin of DNA replication. This might result from collecting the start and end samples at different growth stages – if the cells are rapidly dividing then the area near the origin of replication will be at higher copy number. To remove this effect, for strains on the main chromosome, we computed a smooth estimate of how the fitness of each strain varied with chromosomal position (using the loess function in R) and we subtracted this from the fitness values.

It appears that the median gene in *Z. mobilis* ZM4 has a fitness defect in most conditions. For example, in all of our experiments, the median fitness of genes with annotated functions was below the median fitness of purely hypothetical proteins. This might reflect its relatively small genome (1,892 proteins). Thus, setting the median gene's fitness to zero was not appropriate. Instead, for genes on

the main chromosome, we set the mode of the distribution to zero. (More precisely, we estimated the mode by finding the maximum of the kernel density, using the density function in R with default settings, and we subtracted the mode from the values.) Mode-based centering typically lowered the fitness values by around 0.1. We used mode-based centering for *S. oneidensis* MR-1 and *D. alaskensis* G20 as well, although it made less difference for those organisms.

To identify genes with strong and statistically significant effects on fitness, we used a threshold of ± 0.75. Effects above this magnitude were usually highly statistically significant, with z scores above 3 or below -3. (z scores were computed as described previously: briefly, we used a moderated t-like test statistic for each gene and converted the test statistics to z scores by using control experiments [4].) A fitness of ± 0.75 corresponds to around a 7% change in abundance per generation.

To identify genes with more subtle but reproducible effects on fitness, we grouped together experiments with similar patterns (those having a pairwise correlation of above 0.75). For each group, we used Fisher's method to combine the significance of genes (as assessed using z scores). For each gene, we corrected for multiple testing across groups.

Statistical tools

All statistical analyses were conducted in R 2.11 or 2.13 (http://r-project.org/). Data was visualized in R and in MicrobesOnline [27].

Data availability

All fitness data is available in MicrobesOnline (http://microbesonline.org/). All gene expression, tiling, and 5' RNA-Seq data have been submitted to the Gene Expression Omnibus (GEO). All data and source code are available from the authors' web site (http://genomics.lbl.gov/supplemental/exprVfitness2012/).

Acknowledgements

We thank Dacia Leon, Dan Tarjan, Keith K. Keller, Jason K. Baumohl, and Marcin P. Joachimiak for technical assistance, and Paramvir S. Dehal for helpful discussions. We thank the Energy Biosciences Institute for providing the mutant collection for *Z. mobilis* ZM4.

References

1. Wall ME, Hlavacek WS, Savageau MA (2004) Design of gene circuits: lessons from bacteria. Nat Rev Genet 5: 34–42.

2. Seshasayee AS, Fraser GM, Babu MM, Luscombe NM (2009) Principles of transcriptional regulation and evolution of the metabolic system in E. coli. Genome Res 19: 79–91.

3. Oh J, Fung E, Price M, Dehal P, Davis R, et al. (2010) A universal TagModule collection for parallel genetic analysis of microorganisms. Nucleic Acids Research 38: e146–e146.

4. Deutschbauer A, Price MN, Wetmore KM, Shao W, Baumohl JK, et al. (2011) Evidence-based annotation of gene function in Shewanella oneidensis MR-1 using genome-wide fitness profiling across 121 conditions. PLoS Genet 7: e1002385.

5. Giaever G, Chu AM, Ni L, Connelly C, Riles L, et al. (2002) Functional profiling of the Saccharomyces cerevisiae genome. Nature 418: 387–391.

6. Birrell GW, Brown JA, Wu HI, Giaever G, Chu AM, et al. (2002) Transcriptional response of Saccharomyces cerevisiae to DNA-damaging agents does not identify the genes that protect against these agents. Proc Natl Acad Sci USA 99: 8778–8783.

7. Smith JJ, Sydorskyy Y, Marelli M, Hwang D, Bolouri H, et al. (2006) Expression and functional profiling reveal distinct gene classes involved in fatty acid metabolism. Mol Syst Biol 2: 2006.0009.

8. Fischer E, Sauer U (2005) Large-scale in vivo flux analysis shows rigidity and suboptimal performance of Bacillus subtilis metabolism. Nat Genet 37: 636–40.

9. Wessely F, Bartl M, Guthke R, Li P, Schuster S, et al. (2011) Optimal regulatory strategies for metabolic pathways in Escherichia coli depending on protein costs. Mol Syst Biol 7: 515.

10. Lercher MJ, Pal C (2008) Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Mol Biol Evol 25: 559–567.

11. Price MN, Dehal PS, Arkin AP (2008) Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli. Genome Biol 9: R4.

12. Stone JR, Wray GA, Wray GA (2001) Rapid evolution of cis-regulatory sequences via local point mutations. Mol Biol Evol 18: 1764–1770.

13. Berg J, Willmann S, Lassig M (2004) Adaptive evolution of transcription factor binding sites. BMC Evol Biol 4: 42.

14. Tagkopoulos I, Liu YC, Tavazoie S (2008) Predictive behavior within microbial genetic networks. Science 320: 1313–1317.

15. Mitchell A, Romano GH, Groisman B, Yona A, Dekel E, et al. (2009) Adaptive prediction of environmental changes by microorganisms. Nature 460: 220–224.

16. Lynch M (2007) The evolution of genetic networks by non-adaptive processes. Nature Reviews Genetics 8: 803–813.

17. Rajewsky N, Socci ND, Zapotocky M, Siggia ED (2002) The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. Genome Res 12: 298–308.

18. McCue LA, Thompson W, Carmack CS, Lawrence CE (2002) Factors influencing the identification of transcription factor binding sites by cross-species comparison. Genome Res 12: 1523–32.

19. Price MN, Dehal PS, Arkin AP (2007) Orthologous transcription factors in bacteria have different functions and regulate different genes. PLoS Comput Biol 3: e175.

20. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O (2001) The Comprehensive Microbial Resource. Nucleic Acids Res 29: 123–125.

21. de Daruvar A, Collado-Vides J, Valencia A (2002) Analysis of the cellular functions of Escherichia coli operons and their conservation in Bacillus subtilis. J Mol Evol 55: 211–21.

22. Rogozin IB, Makarova KS, Murvai J, Czabarka E, Wolf YI, et al. (2002) Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res 30: 2212–23.

23. Price MN, Arkin AP, Alm EJ (2006) The life-cycle of operons. PLoS Genet 2: e96.

24. Langridge G, Phan M, Turner D, Perkins T, Parts L, et al. (2009) Simultaneous assay of every salmonella typhi gene using one million transposon mutants. Genome research 19: 2308–2316.

25. Koskiniemi S, Sun S, Berg O, Andersson D (2012) Selection-driven genome reduction in bacteria. PloS Genetics .

26. Price MN, Huang KH, Alm EJ, Arkin AP (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic Acids Res 33: 880–92.

27. Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, et al. (2009) Microbesonline: an integrated portal for comparative and functional genomics. Nucleic Acids Res database issue.

28. Novichkov PS, Laikova ON, Novichkova ES, Gelfand MS, Arkin AP, et al. (2010) Regprecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. Nucleic Acids Res Database issue: D111-8.

29. Prouty M, Correa N, Klose K (2001) The novel σ54-and σ28-dependent flagellar gene transcription hierarchy of Vibrio cholerae. Molecular microbiology 39: 1595–1609.

30. Dasgupta N, Wolfgang M, Goodman A, Arora S, Jyot J, et al. (2003) A four-tiered transcriptional regulatory circuit controls flagellar biogenesis in Pseudomonas aeruginosa. Molecular microbiology 50: 809–824.

31. Bremer H, Dennnis PP (1996) Modulation of chemical composition and other parameters of the cell by growth rate. In: Escherichia coli and Salmonella typhimurium: cellular and molecular biology., American Society for Microbiology. 2 edition, pp. 1553-1569.

32. Lemke JJ, Sanchez-Vazquez P, Burgos HL, Hedberg G, Ross W, et al. (2011) Direct regulation of Escherichia coli ribosomal protein promoters by the transcription factors ppGpp and DksA. Proc Natl Acad Sci USA 108: 5712–5717.

33. Paul B, Barker M, Ross W, Schneider D, Webb C, et al. (2004) DksA: a critical component of the transcription initiation machinery that potentiates the regulation of rRNA promoters by ppGpp and the initiating NTP. Cell 118: 311–322.

34. Travers A (1980) Promoter sequence for stringent control of bacterial ribonucleic acid synthesis. Journal of bacteriology 141: 973–976.

35. Haugen S, Berkmen M, Ross W, Gaal T, Ward C, et al. (2006) rrna promoter regulation by nonoptimal binding of σ region 1.2: An additional recognition element for rna polymerase. Cell 125: 1069–1082.

36. Saffarini D, Schultz R, Beliaev A (2003) Involvement of cyclic amp (cAMP) and cAMP receptor protein in anaerobic respiration of Shewanella oneidensis. Journal of bacteriology 185: 3668–3671.

37. Cruz-García C, Murray AE, Rodrigues JL, Gralnick JA, McCue LA, et al. (2011) Fnr (EtrA) acts as a fine-tuning regulator of anaerobic metabolism in Shewanella oneidensis MR-1. BMC Microbiology 11: 64.

38. Serres M, Riley M (2006) Genomic analysis of carbon source metabolism of Shewanella oneidensis MR-1: predictions versus experiments. Journal of bacteriology 188: 4601–4609.

39. Pinchuk G, Ammons C, Culley D, Li S, McLean J, et al. (2008) Utilization of DNA as a sole source of phosphorus, carbon, and energy by Shewanella spp.: ecological and physiological implications for dissimilatory metal reduction. Applied and environmental microbiology 74: 1198–1208.

40. Pinchuk G, Hill E, Geydebrekht O, De Ingeniis J, Zhang X, et al. (2010) Constraint-based model of Shewanella oneidensis MR-1 metabolism: a tool for data analysis and hypothesis generation. PLoS computational biology 6: e1000822.

41. Charoensawan V, Wilson D, Teichmann SA (2010) Genomic repertoires of DNA-binding transcription factors across the tree of life. Nucleic Acids Res 38: 7364–7377.

42. Martinez-Antonio A, Collado-Vides J (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. Curr Opin Microbiol 6: 482–489.

43. Sengupta A, Djordjevic M, Shraiman B (2002) Specificity and robustness in transcription control networks. Proceedings of the National Academy of Sciences 99: 2072.

44. Lozada-Chávez I, Angarica V, Collado-Vides J, Contreras-Moreira B (2008) The role of DNA-binding specificity in the evolution of bacterial regulatory networks. Journal of molecular biology 379: 627–643.

45. Chubukov V, Zuleta IA, Li H (2012) Regulatory architecture determines optimal regulation of gene expression in metabolic pathways. Proc Natl Acad Sci USA 109: 5127–5132.

46. Conrad TM, Lewis NE, Palsson B (2011) Microbial laboratory evolution in the era of genome-scale science. Mol Syst Biol 7: 509.

47. Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, et al. (2010) The primary transcriptome of the major human pathogen Helicobacter pylori. Nature 464: 250–255.

48. Wolf Y, Rogozin IB, Kondrashov AS, Koonin EV (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res 11: 356–72.

49. Ermolaeva MD, White O, Salzberg SL (2001) Prediction of operons in microbial genomes. Nucleic Acids Res 29: 1216–21.

50. Itoh T, Takemoto K, Mori H, Gojobori T (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. Mol Biol Evol 16: 332–46.

51. van Nimwegen E (2003) Scaling laws in the functional content of genomes. Trends Genet 19: 479–484.

52. Itzkovitz S, Tlusty T, Alon U (2006) Coding limits on the number of transcription factors. BMC Genomics 7: 239.

53. Sasson V, Shachrai I, Bren A, Dekel E, Alon U (2012) Mode of regulation and the insulation of bacterial gene expression. Molecular Cell 46: 399–407.

54. Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muñiz-Rascado L, et al. (2011) RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (gensor units). Nucleic Acids Research 39: D98–D105.

55. Pierce S, Davis R, Nislow C, Giaever G (2007) Genome-wide analysis of barcoded saccharomyces cerevisiae gene-deletion mutants in pooled cultures. Nature protocols 2: 2958–2974.

56. Liu Y, Gao W, Wang Y, Wu L, Liu X, et al. (2005) Transcriptome analysis of Shewanella oneidensis MR-1 in response to elevated salt conditions. J Bacteriol 187: 2501-7.

57. Faith J, Driscoll M, Fusaro V, Cosgrove E, Hayete B, et al. (2008) Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata. Nucleic acids research 36: D866–D870.

58. Price MN, Deutschbauer AM, Kuehl JV, Liu H, Witkowska HE, et al. (2011) Evidence-based annotation of transcripts and proteins in the sulfate-reducing bacterium Desulfovibrio vulgaris Hildenborough. J Bacteriol 193: 5716–5727.

59. Güell M, van Noort V, Yus E, Chen W, Leigh-Bell J, et al. (2009) Transcriptome complexity in a genome-reduced bacterium. Science 326: 1268-71.

60. Goodman A, Rogers P, Skotnicki M (1982) Minimal medium for isolation of auxotrophic Zymomonas mutants. Applied and environmental microbiology 44: 496.

61. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, et al. (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2: 2006.0008.

62. Allen TE, Herrgard MJ, Liu M, Qiu Y, Glasner JD, et al. (2003) Genome-scale analysis of the uses of the Escherichia coli genome: model-driven analysis of heterogeneous data sets. J Bacteriol 185: 6392–9.

Financial Disclosure

This work conducted by ENIGMA was supported by the Office of Science, Office of Biological and Environmental Research, of the U. S. Department of Energy under Contract No. DE-AC02-05CH11231. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abbreviations

LB: Luria-Bertani broth. NAG: N-acetylglucosamine.

Supporting Information

Table S1 – Pairs of functionally-related genes in Shewanella oneidensis MR-1 that are not in the same operon and are not coexpressed

We list pairs of genes that are cofit and in the same functional category (TIGR subrole) but are not in the same operon, near each other in the genome, or coexpressed. For each pair, we manually examined their annotations and their fitness patterns to determine if they truly had closely-related functions or not. For pairs of flagellar genes, we also report whether they are coregulated and in the same "class" in *Pseudomonas aeruginosa* according to Dasgupta *et al.* 2003.

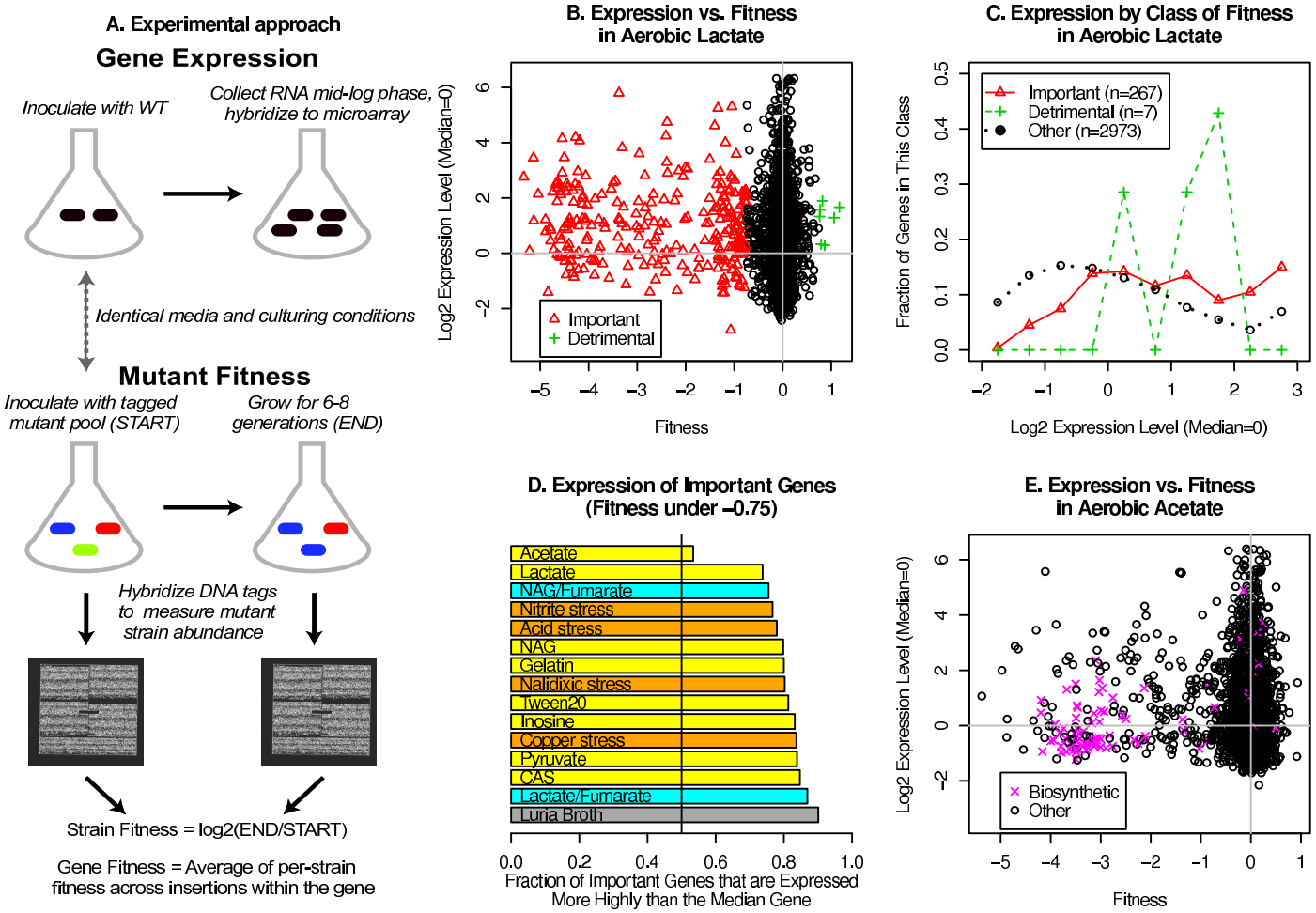

Figure 1 - Mutant fitness versus absolute gene expression levels in Shewanella oneidensis MR-1. (A) Our experimental approach. Each pool of mutants contains about 4,000 strains, and each strain has a transposon inserted at a different location in the genome and a tag that allows that strain to be distinguished from the other strains in that pool [3]. (B) Expression level versus fitness during aerobic growth in minimal lactate medium. Genes with fitness values below -0.75 (important for fitness) or above 0.75 (detrimental to fitness) are color-coded. (C) Another view of the expression levels from panel B: the distribution of absolute expression for genes that are important for fitness, detrimental to fitness, or have little phenotype. Expression values beyond the range of the plot are included in the left-most or right-most bins. (D) In each of 15 diverse conditions, the fraction of important genes that are expressed more highly than the median gene. The vertical line shows the random expectation of 0.5. Conditions are color-coded by type: yellow for carbon sources, orange for stresses, cyan for anaerobic, and grey for LB. (E) Expression level versus fitness during aerobic growth in minimal acetate medium. Biosynthetic genes, as annotated by TIGRFam [20], are highlighted.

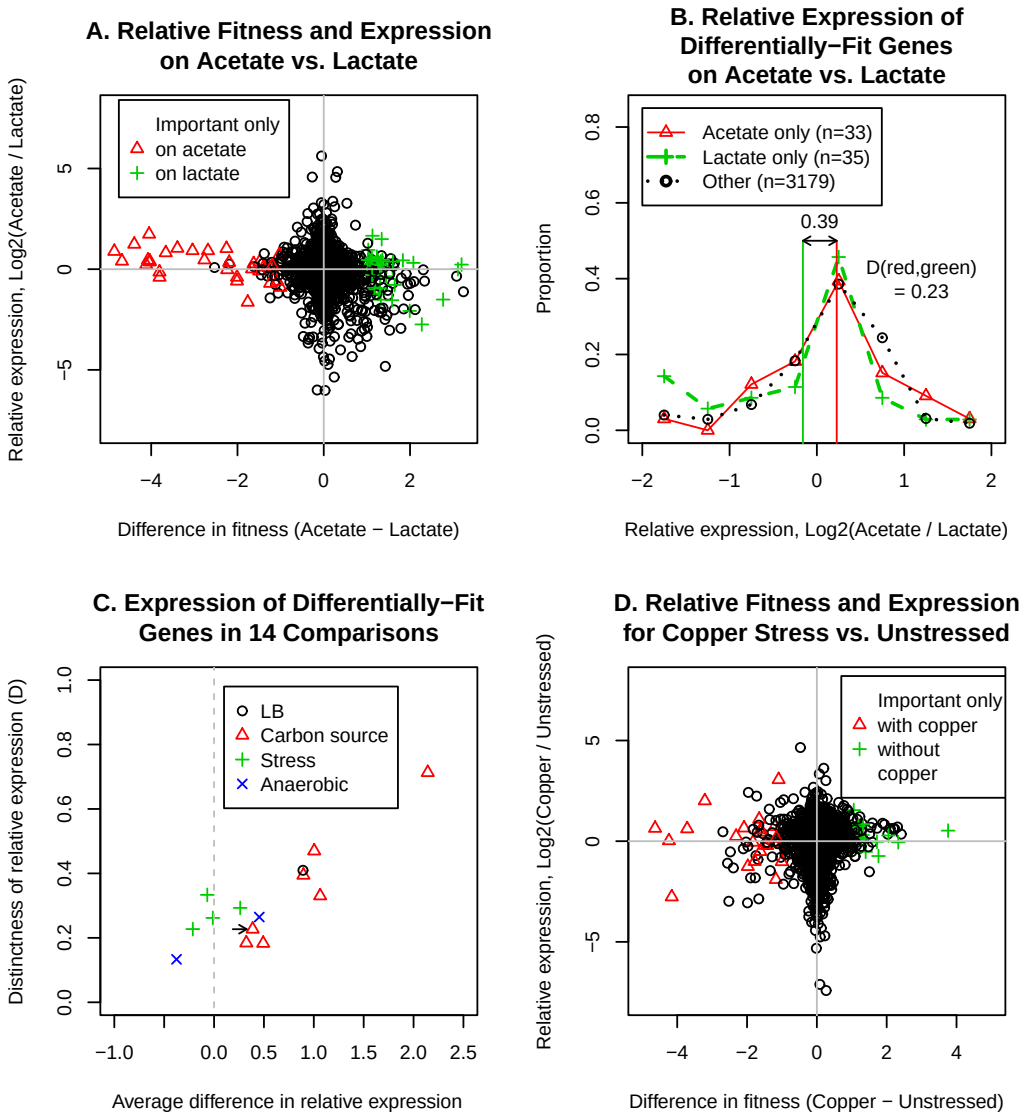

Figure 2 - Differential fitness and relative expression in Shewanella oneidensis MR-1. (A) Relative expression versus the difference in fitness for aerobic growth on acetate versus aerobic growth on lactate. Genes are color-coded if they are important for fitness on acetate or lactate but not the other condition (specifically, if fitness is below -0.75 in that condition but not in the other condition and if the difference in fitness between the conditions is at least 1.0). (B) Another view of the relative expression from panel A: the distribution of relative expression for genes that are only important on acetate, only important on lactate, or other genes. Out-of-range values are included in the left- or right-most bins. The vertical lines show the averages for genes that are important only in acetate (green line) or only in lactate (red line). (C) The change in expression of differentially-fit genes in each of 14 conditions when compared to aerobic lactate. The x axis shows the average difference in relative expression between the two groups of genes, while the y axis shows the Kolmogorov-Smirnov D statistic for how distinct the two distributions are. The arrow highlights the comparison between acetate and lactate from panel B. (D) Relative expression versus the difference in fitness for cells growing in minimal lactate medium with or without copper added.

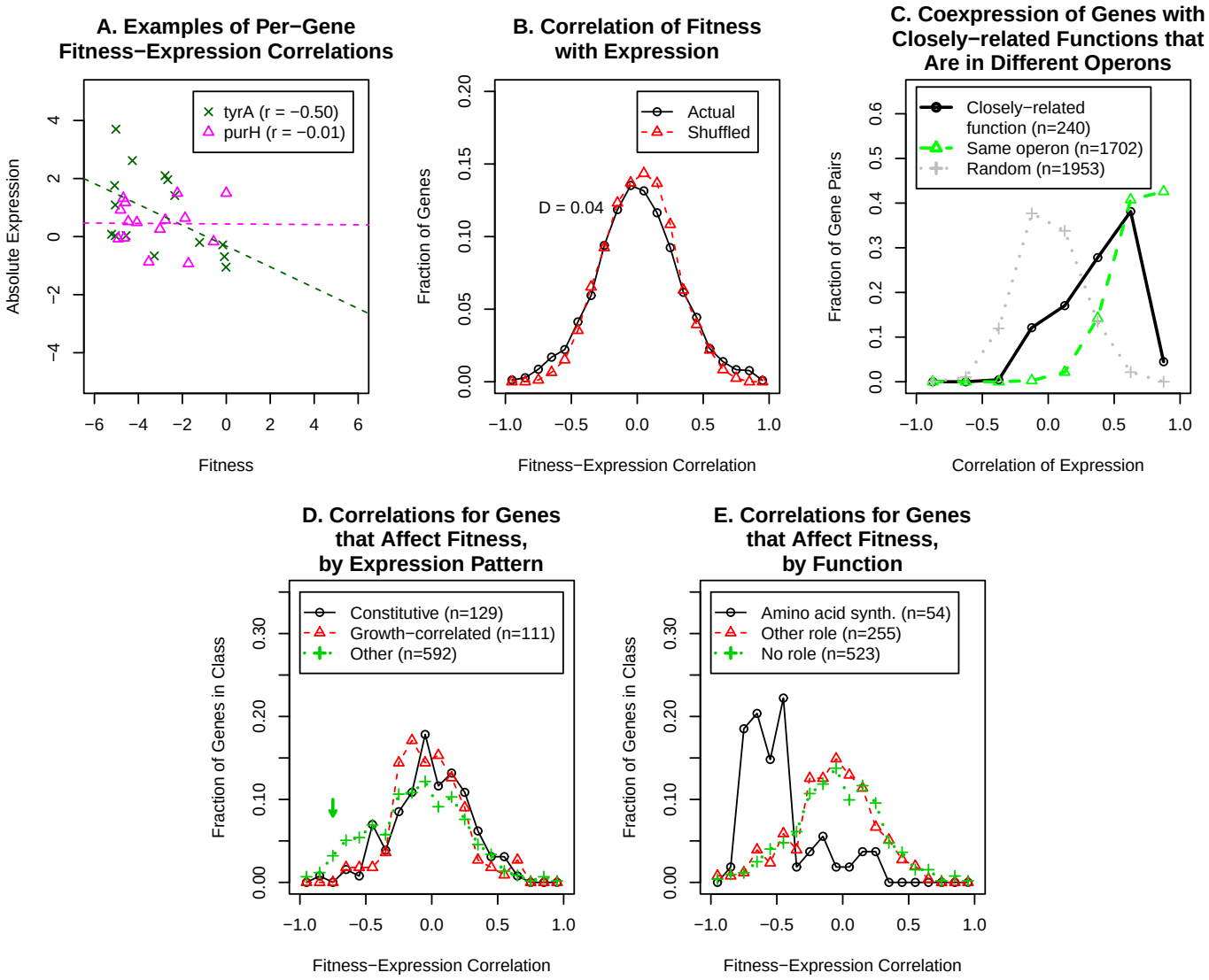

Figure 3 - The correlation of fitness and expression in Shewanella oneidensis MR-1. (A) Absolute expression versus fitness for *tyrA* and *purH* across 15 growth conditions. The lines show the best fit for each gene: *tyrA* tends to be expressed more highly when it is more important for fitness (r = -0.50), but *purH* does not (r = -0.01). (B) The distribution of fitness-expression correlations, computed as in panel A, for 3,247 genes and for 3,247 shuffled controls. (C) The coexpression, across 329 experiments, of pairs of genes that are not in the same operon and have closely-related functions (i.e., matching TIGR subroles and similar patterns of mutant fitness across 195 experiments). We also show the distribution of coexpression for genes that are predicted to be in the same operon and for random pairs of genes that have different TIGR subroles and are not adjacent or predicted to be in the same operon. (D & E) The distribution of fitness-expression correlations (as in panel B) when considering only genes that have fitness of above 0.75 or below -0.75 in at least one of the 15 conditions. In (D), we separate out constitutive and growth-correlated genes from other genes, and the arrow highlights the adaptive regulation of some of the other genes. In (E), the genes are classified by their TIGR roles.

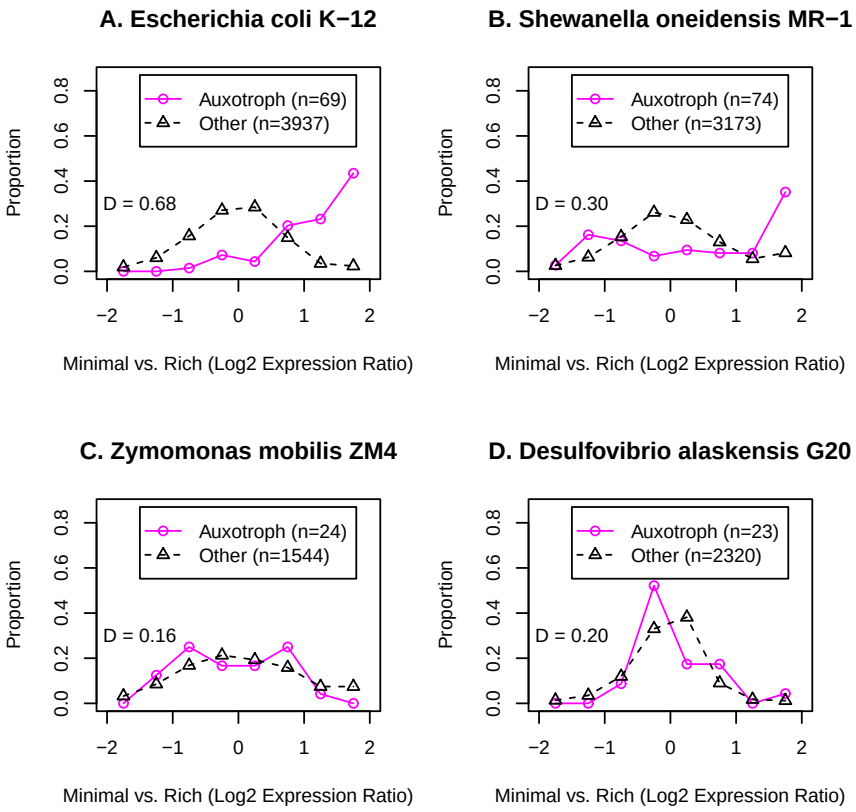

Figure 4 - The relative expression of biosynthetic genes in minimal versus rich media across diverse bacteria. We examined whether auxotrophs were upregulated in minimal media, as compared to other genes, in (A) *Escherichia coli* K-12; (B) *Shewanella oneidensis* MR-1; (C) *Zymomonas mobilis* ZM4; and (D) *Desulfovibrio alaskensis* G20. In all four organisms, the auxotrophs are annotated by TIGR role as being involved in amino acid, nucleotide, or cofactor synthesis, and experimental data confirms that they are important for growth in a defined medium but not in rich medium. For *E. coli* K-12, we used growth data of deletion mutants from the Keio collection [61] and expression data from [62]. For the other organisms, we collected fitness data using pooled transposon mutants and we collected gene expression data using microarrays. Genes were considered important only in defined medium if their fitness was below -0.75 in defined medium but not in rich medium and the difference in fitness was at least 1. The expression $\log_2$ ratios are normalized so that the median value is 0. Log_2 ratios that are below -2 or above 2 are included in the left- or right-most bins, respectively.